

Chapter 6

Emotion Modeling

Emotions are a key part of how humans think, make decisions, and interact with others. They guide us to understand situations, make choices, and build relationships. Antonio Damasio, in his book *Descartes' Error* [25], explained that emotions are not separate from logic. Instead, they are deeply connected to how we reason and act. When developing LLM agents, adding emotional capabilities can potentially make these systems smarter, more adaptable, and better understand the world around them.

For LLM agents, emotions can act as a decision-making tool, much like they do for humans. Emotions help us prioritize tasks, understand risks, and adapt to new challenges. Marvin Minsky, in *The Emotion Machine* [420], described emotions as a way to adjust our thinking processes, helping us solve problems in a more flexible and creative manner. Similarly, LLM agents with emotion-like features could improve their ability of solving complex problems and making decisions in a more human-style.

However, the integration of emotions into LLM agents is still in its early stages. Researchers are just starting to explore how emotional capabilities can improve these systems. Furthermore, there is great potential for LLM agents to support human emotional well-being, whether through empathetic conversations, mental health support, or simply building better connections with users. This promising but challenging area requires collaboration between fields such as psychology, cognitive science, and AI ethics. As research advances, emotion-understanding LLM agents could redefine how we interact with technology, creating deeper trust and more meaningful relationships between humans and machines.

In the following subsections, we will delve deeper into the role of emotions in shaping LLM agents. We will explore how emotions can be used to enhance learning and adaptability, how LLMs understand human emotions, and how these systems express and model their own emotional states. We will also examine how emotions can be manipulated to influence LLM agents' behavior and personalities, as well as the ethical and safety concerns that arise from these capabilities. Each of these discussions builds on the foundational importance of emotion to create LLM agents that are more intelligent, empathetic, and aligned with human values.

6.1 Psychological Foundations of Emotion

Psychological and neuroscientific theories of emotion provide essential frameworks for developing emotionally intelligent LLM agents. These theories can be categorized into several major approaches, each offering unique perspectives on how emotions function and how they might be implemented in AI systems.

Categorical Theories. These models posit that emotions exist as discrete, universal categories with distinct physiological and behavioral signatures. Ekman's theory of basic emotions [421] identifies six fundamental emotions (anger, disgust, fear, happiness, sadness, and surprise) that are recognized across cultures and expressed through specific facial configurations. This discrete approach has significantly influenced affective computing, with many emotion classification systems in AI adopting these labels for training [422, 423]. For LLM agents, categorical frameworks provide clear taxonomies for classifying user emotions and generating appropriate responses. However, they face criticism for oversimplifying the complex, blended nature of human emotional experience [424] and may not capture cultural variations in emotional expression [425].

Dimensional Models. Rather than discrete categories, dimensional approaches represent emotions as points in a continuous space defined by fundamental dimensions. Russell’s Circumplex Model [426] maps emotions onto two primary dimensions: valence (pleasure-displeasure) and arousal (activation-deactivation). This framework enables more nuanced tracking of emotional states. It distinguishes between high-arousal panic and low-arousal anxiety despite both having negative valence. The PAD (Pleasure-Arousal-Dominance) model [427] extends this by adding a dominance dimension, capturing the sense of control or power associated with emotional states. These continuous representations have proven valuable for LLM systems that need to generate emotionally graded responses or track subtle shifts in user affect over time [428, 429, 430]. Dimensional models allow for fine-grained control over generated content, enabling humans or agents to modulate tone along continuous scales rather than switching between discrete emotional states.

Hybrid and Componential Frameworks. Recognizing limitations in pure categorical or dimensional approaches, several theories integrate aspects of both. Plutchik’s Wheel of Emotions [431] arranges eight primary emotions in a wheel structure with intensity gradients and dimensional properties, allowing for the representation of complex emotional blends (e.g., love as a mixture of joy and trust). Meanwhile, componential models like Scherer’s Component Process Model (CPM) [432] conceptualize emotions as emerging from synchronized components including cognitive appraisal, physiological arousal, action tendencies, and subjective feelings. Particularly influential in AI research is the OCC (Ortony-Clore-Collins) model [433], which defines 22 emotion types based on how events, agents, or objects are evaluated relative to goals and standards. These appraisal-based frameworks have been implemented in dialogue systems that generate emotional responses through rule-based evaluation of situations [434, 435]. For LLM agents, such models provide computational structures for evaluating text input and selecting contextually appropriate emotional responses, improving both coherence and perceived empathy [436, 437].

Neurocognitive Perspectives. The neuroscience of emotion offers additional insights for LLM architectures. Damasio’s somatic marker hypothesis [25] emphasizes how emotions, implemented through body-brain interactions, guide decision-making by associating physiological states with anticipated outcomes. This interaction between the limbic system and the cortex shows a two-process architecture: fast “alarm” signals in the limbic system, like those processed by the amygdala, work alongside slower, more deliberate reasoning in the cortex. Contemporary LLM systems have begun implementing analogous architectures, where fast sentiment detection modules work in parallel with more thorough chain-of-thought reasoning [436, 437]. Recent evidence further suggests that opponent circuitry in the striatum enables distributional reinforcement learning by encoding not just mean rewards but entire probability distributions, offering a neural basis for emotion-influenced decision-making under uncertainty [438]. Similarly, LeDoux’s distinction between “low road” (quick, automatic) and “high road” (slower, cognitive) fear processing [24] suggests design patterns for systems that need both immediate safety responses and nuanced emotional understanding. Minsky’s framing of emotions as “ways to think” [420] that reorganize cognitive processes has influenced frameworks like EmotionPrompt [428] and Emotion-LLaMA [423], where emotional context dynamically reshapes LLM reasoning.

These theoretical frameworks increasingly inform the development of emotionally intelligent LLM agents. Categorical models provide clear labels for emotion classification tasks [423, 429], while dimensional embeddings enable continuous control over generated text [428]. Hybrid approaches help systems handle mixed emotions and emotional intensity. Appraisal-based methods, particularly those derived from the OCC model, allow LLMs to evaluate narrative events or user statements contextually, selecting appropriate emotional responses that foster rapport and trust [439]. Neuroscientifically-inspired dual-process architectures combine “fast” sentiment detection with “slow” deliberative reasoning, enabling both quick safety responses and deeper emotional understanding [436, 437]. While explicit neurocognitive mechanisms (like dedicated “amygdala-like” pathways) remain rare in current LLM pipelines, emerging research explores biologically-inspired modules to handle urgent emotional signals and maintain consistent emotional states across extended interactions [440, 441].

Emotion is a key part of human intelligence, and it will likely become one of the key components or design considerations of LLM agents. One key future direction is systematically translating these psychological and neuroscience theories into an LLM agent’s internal processes. Techniques for translating might include using dimensional models (e.g., valence/arousal/dominance) as latent states that influence generation or adopting explicit rule-based appraisals (OCC) to label user messages and shape the agent’s subsequent moves. Hybrid approaches offer a compelling balance: an LLM could first recognize a discrete category (e.g., “fear”) but also gauge its intensity and control dimension for finer-grained conversation. Such emotion-infused architectures might yield more coherent “moods” over time, analogous to how humans sustain affective states rather than resetting at every turn. Explicit alignment with psychological theories also enhances interpretability: designers can debug or refine the agent’s responses by comparing them to well-established emotion constructs, rather than dealing with opaque emergent behaviors.

A second direction is harnessing these theories to improve *affectionate or supportive interactions*, often referred to as emotional alignment. For example, circumplex or PAD-based tracking can help an LLM detect negative valence and high arousal in a user’s text and respond soothingly (e.g., lowering arousal, offering empathetic reappraisals). In

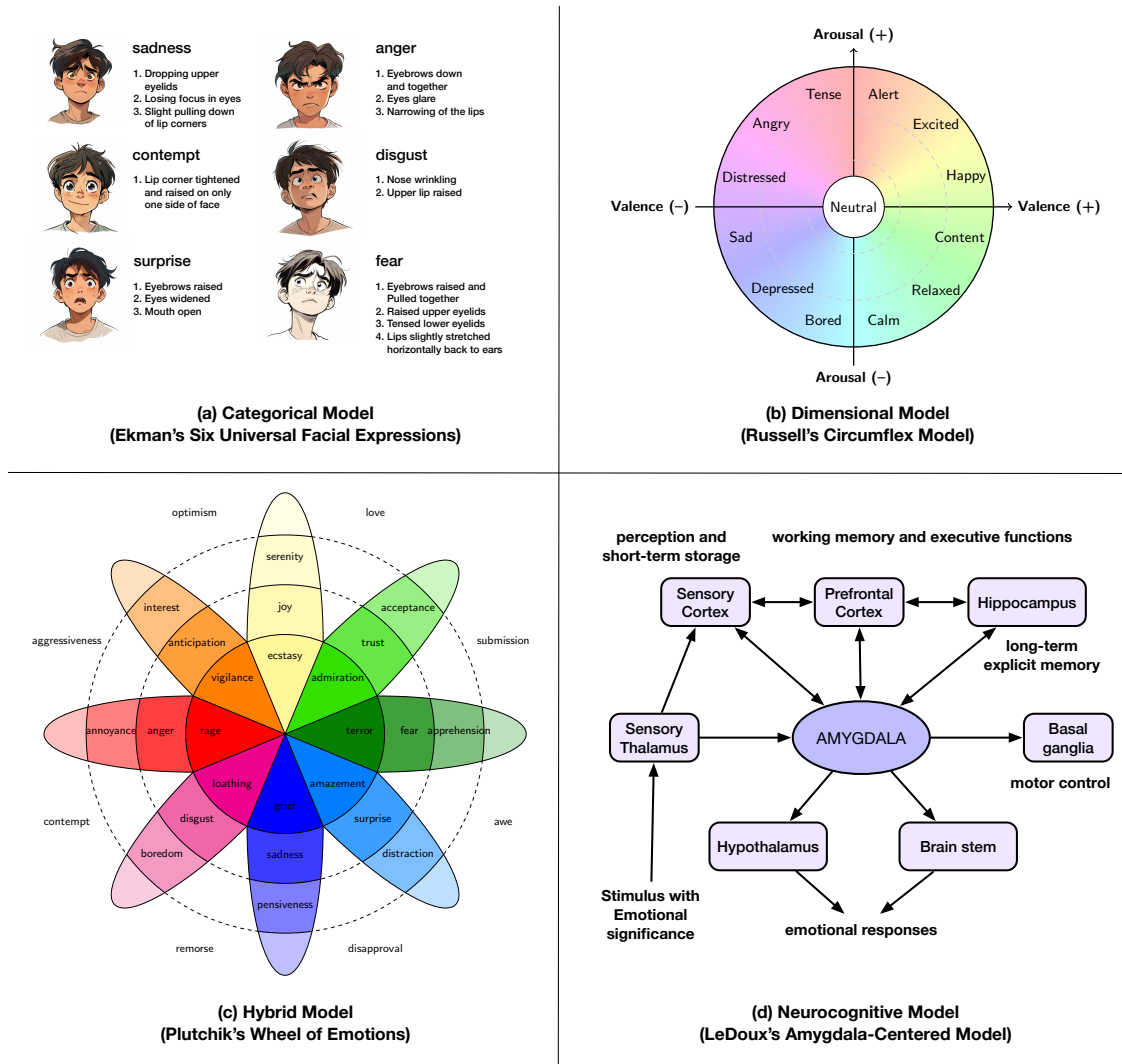


Figure 6.1: Visualization and examples of major emotion theory categories. (a) Categorical Theories: Ekman's six basic emotions [421] showing discrete emotional states. (b) Dimensional Models: Russell's Circumplex [426] representing emotions as coordinates in continuous space. (c) Hybrid/Componential Frameworks: Plutchik's Wheel [431] combining intensity gradients with categorical emotions. (d) Neurocognitive Perspectives: LeDoux's Amygdala-Centered Model [24] showing dual-pathway processing of emotional stimuli. These psychological foundations inform different approaches to emotion modeling in AI systems, from discrete classification to dimensional representations, appraisal-based reasoning, and multi-pathway information processing.

mental health or counseling scenarios, an appraisal-informed method could let the agent validate the user's feelings and understand their situation in terms of goal incongruence or perceived blame, which helps craft responses that convey genuine empathy. Grounding emotional outputs in cognitive theories (like "relief" if a negative outcome is avoided, or "gratitude" when a user helps the system) likewise makes interactions feel more natural and ethically aligned. These enhancements are particularly salient as LLMs migrate into real-world applications like customer service, elder care, and tutoring, where emotional sensitivity can improve outcomes and user well-being. By incorporating robust psychological and limbic-system insights, developers can design LLM agents that not only reason more effectively but also provide sincere emotional support, bridging the gap between computational precision and human-centric care.

6.2 Incorporating Emotions in AI Agents

The integration of emotional intelligence into large language models (LLMs) has emerged as a transformative approach to enhancing their performance and adaptability. Recent studies, such as those of EmotionPrompt [422], highlight how emotional stimuli embedded in prompts can significantly improve outcomes across various tasks, including a notable 10.9% improvement in generative task metrics such as truthfulness and responsibility. By influencing the attention mechanisms of LLMs, emotionally enriched prompts enrich representation layers and result in more nuanced outputs [422]. These advancements bridge AI with emotional intelligence, offering a foundation for training paradigms that better simulate human cognition and decision-making, particularly in contexts requiring social reasoning and empathy.

Multimodal approaches further elevate the impact of emotional integration. Models like Emotion-LLaMA [440] demonstrate how combining audio, visual, and textual data enables better recognition and reasoning of emotions. Using datasets such as MERR [440], these models align multimodal inputs into shared representations, facilitating improved emotional understanding and generation. This innovation extends beyond linguistic improvements, offering applications in human-computer interaction and adaptive learning. Together, these methods underscore the critical role of emotions in bridging technical robustness with human-centric AI development, paving the way for systems that are both intelligent and empathetic.

6.3 Understanding Human Emotions through AI

Textual Approaches. Recent work highlights the ability of LLMs to perform detailed reasoning about latent sentiment and emotion. Using step-by-step prompting strategies, such as chain of thought reasoning, researchers enable LLMs to infer sentiment even when explicit cues are absent [436]. Beyond single-turn inference, negotiation-based frameworks further refine emotional judgments by leveraging multiple LLMs that cross-evaluate each other's outputs, effectively mimicking a more deliberative human reasoning process [437]. These techniques underscore the importance of iterative, context-aware strategies to capture subtle emotional signals from purely textual input.

Multimodal Approaches. LLMs have also been extended to integrate signals from audio, video, and images. Recent efforts show how additional contextual or world knowledge can be fused with visual and textual information to capture deeper affective states [442]. Moreover, frameworks that convert speech signals into textual prompts demonstrate that vocal nuances can be embedded in LLM reasoning without changing the underlying model architecture [443]. This multimodal integration, combined with explainable approaches, allows for richer and more transparent representations of emotional content [444].

Specialized Frameworks. Beyond generic techniques, specialized systems address tasks in which emotion recognition requires higher levels of awareness of ambiguity [439], context sensitivity, and generative adaptability [445]. These approaches emphasize the inherent complexity of human emotion, treating it as dynamic and probabilistic rather than strictly categorical. Using flexible LLM instruction paradigms, they offer pathways to better interpret ambiguous emotional expressions and integrate contextual cues (e.g., dialogue history), moving LLM closer to human-like emotional comprehension.

Evaluation and Benchmarks. To holistically assess the emotional intelligence of LLM, researchers have proposed various benchmark suites. Some focus on generalized emotion recognition across different modalities and social contexts [446, 447], while others compare the performance and efficiency of models of varying sizes [448]. There are also specialized benchmarks that evaluate multilingual capabilities [449], annotation quality [450], or empathetic dialogue systems [451]. Furthermore, frameworks such as EMOBENCH [441] and MEMO-Bench [452] test nuanced emotional understanding and expression in both text and images, while MERBench [453] and wide-scale evaluations [454] address standardization concerns in multimodal emotion recognition. Together, these benchmarks reveal the growing, yet still imperfect grasp of human emotion by LLMs, highlighting ongoing challenges such as implicit sentiment detection, cultural adaptation, and context-dependent empathy [455].

6.4 Analyzing AI Emotions and Personality

Reliability of Personality Scales for LLMs. Large language models (LLMs) show conflicting evidence when evaluated through human-centered personality tests. On one hand, some studies challenge the validity of common metrics, reporting biases such as “agree bias” and inconsistent factor structures, raising doubts about whether these instruments capture genuine traits [456, 457]. On the other hand, systematic experiments reveal that LLMs can exhibit stable, human-like trait patterns and even adapt to different personas under specific prompts [458, 459]. Yet, concerns persist

about action consistency, alignment of self-knowledge, and whether role-playing agents truly maintain fidelity to their assigned characters [460, 461].

Psychometric Methods & Cognitive Modeling Approaches. Recent work applies rigorous psychometric testing, cognitive tasks, and population-based analyses to uncover how LLM processes and represents mental constructs [462, 463, 464]. Fine-tuning on human behavioral data can align models with decision patterns that mirror individual-level cognition, while population-based sampling techniques expose variability in neural responses [465, 466]. By merging psychological theories with advanced prompting and embedding methods, researchers illuminate latent representations of constructs like anxiety or risk-taking, showing how LLMs can approximate human reasoning across tasks.

Emotion Modeling. Studies on LLM-based emotional intelligence reveal notable abilities to interpret nuanced affect and predict emotion-laden outcomes, often surpassing average human baselines in standard tests [423, 429]. However, these models do not necessarily emulate human-like emotional processes; they rely on high-dimensional pattern matching that sometimes fails under changing contexts, negative input, or conflicting cues [467, 468]. However, hierarchical emotion structures, coping strategies, and empathy-like behaviors can emerge in larger-scale models, underscoring both the promise of emotional alignment and the ethical challenges in creating AI systems that appear and occasionally function as affective agents.

6.5 Manipulating AI Emotional Responses

Prompt-based Methods. Recent research shows that adopting specific personas or roles through well-engineered prompts can bias LLM cognition, allowing targeted emotional or personality outcomes [469, 470, 471, 472]. By inserting instructions such as “If you were a [persona]”, LLMs adapt not only their thematic style, but also their underlying emotional stance. This approach is powerful for real-time manipulation, though it can be inconsistent across tasks and model variants, highlighting the need for more systematic methods.

Training-based Methods. Fine-tuning and parameter-efficient strategies offer deeper, more stable ways to induce or alter LLM emotions [473, 428, 474]. Quantized Low-Rank Adaptation (QLoRA) and specialized datasets can embed nuanced traits such as the Big Five or MBTI profiles directly into the model’s learned weights. These methods enable LLMs to spontaneously exhibit trait-specific behaviors (including emoji use) and sustain their emotional states over longer dialogues, while also offering interpretability through neuron-level activation patterns.

Neuron-based Methods. A recent advance isolates personality-specific neurons and manipulates them directly to evoke or suppress emotional traits [475]. By toggling neuron activations pinpointed through psychologically grounded benchmarks (e.g., PersonalityBench), LLMs can embody targeted emotional dimensions without retraining the entire network. This neuron-centric approach provides fine-grained, dynamic control over model behaviors, representing a leap in precision and efficiency for emotional manipulation in LLMs.

6.6 Summary and Discussion

Manipulation and Privacy Concerns. The rapid adoption of Emotional AI in advertising and politics raises significant manipulation and privacy risks [476, 477]. Emotional AI often collects sensitive biometric data, such as facial expressions and voice tones, to infer emotional states, enabling targeted advertising or political influence. However, these systems can exploit human emotions for profit or political gain, infringing on fundamental rights and fostering over-surveillance in public spaces [478, 477]. Regulatory frameworks like GDPR and the EU AI Act are critical to mitigating these risks responsibly.

Alignment Issues. Emotional AI’s capacity to detect and interpret emotions is often misaligned with intended outcomes, leading to inaccuracies and biases. Anxiety-inducing prompts, for instance, have been shown to exacerbate biases in large language models (LLMs), affecting outputs in high-stakes domains such as healthcare and education [479, 480]. Misinterpretation of emotional cues by AI systems, as seen in workplace applications, can exacerbate discrimination and power imbalances [481]. Techniques like reinforcement learning from human feedback (RLHF) have proven effective in mitigating these issues but require further development to ensure robust alignment in diverse contexts [479, 423].

Ethical Implications. Trust and acceptance of AI systems are significantly influenced by their ability to exhibit empathy and maintain socially appropriate behavior [482, 483]. However, the commodification of emotions in workplace management and customer service has raised concerns about ethical labor practices and AI-human relationships [481]. Moreover, Emotional AI’s reliance on anthropomorphic characteristics without sufficient empathy can undermine user trust [482]. Frameworks like SafeguardGPT, which incorporate psychotherapy techniques, demonstrate promising approaches to fostering trust and aligning AI behavior with societal norms [484]. Nonetheless, challenges remain in ensuring privacy, fairness, and cultural sensitivity [484, 483].

Distinguishing AI Emotional Mimicry from Human Experience. Despite advances in emotion modeling for LLM agents, a fundamental distinction remains: these systems do not actually “feel” emotions as humans do but only show human-emotion-like patterns via probabilistic modeling. While LLMs can convincingly simulate emotional responses, recognize emotional patterns, and generate affectional outputs, they lack the embodied, phenomenological experience that defines human emotions. This simulation-reality gap creates both technical and ethical challenges. Users frequently anthropomorphize AI systems that display emotion-like behaviors [482], potentially leading to misplaced trust or expectations. This distinction needs to be carefully thought in both research and deployment contexts, as the perceived emotional capabilities of LLMs influence human-AI relationships, ethical frameworks, and regulatory approaches. Future work should balance enhancing LLMs’ emotional intelligence while maintaining transparency about their fundamental limitations as non-sentient systems.